**腾讯视频爬虫采集说明文档**

21大数据技术1班 肖宇宁 2104230110

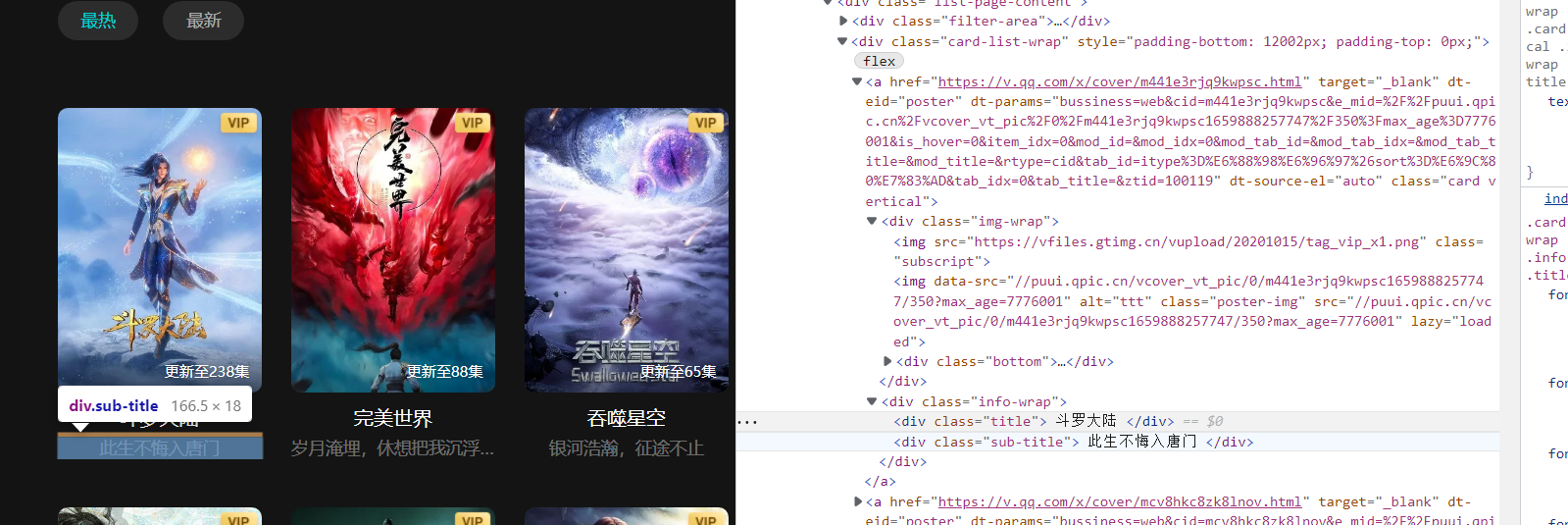
一.数据采集目标

本次爬虫采集围绕以下内容展开：

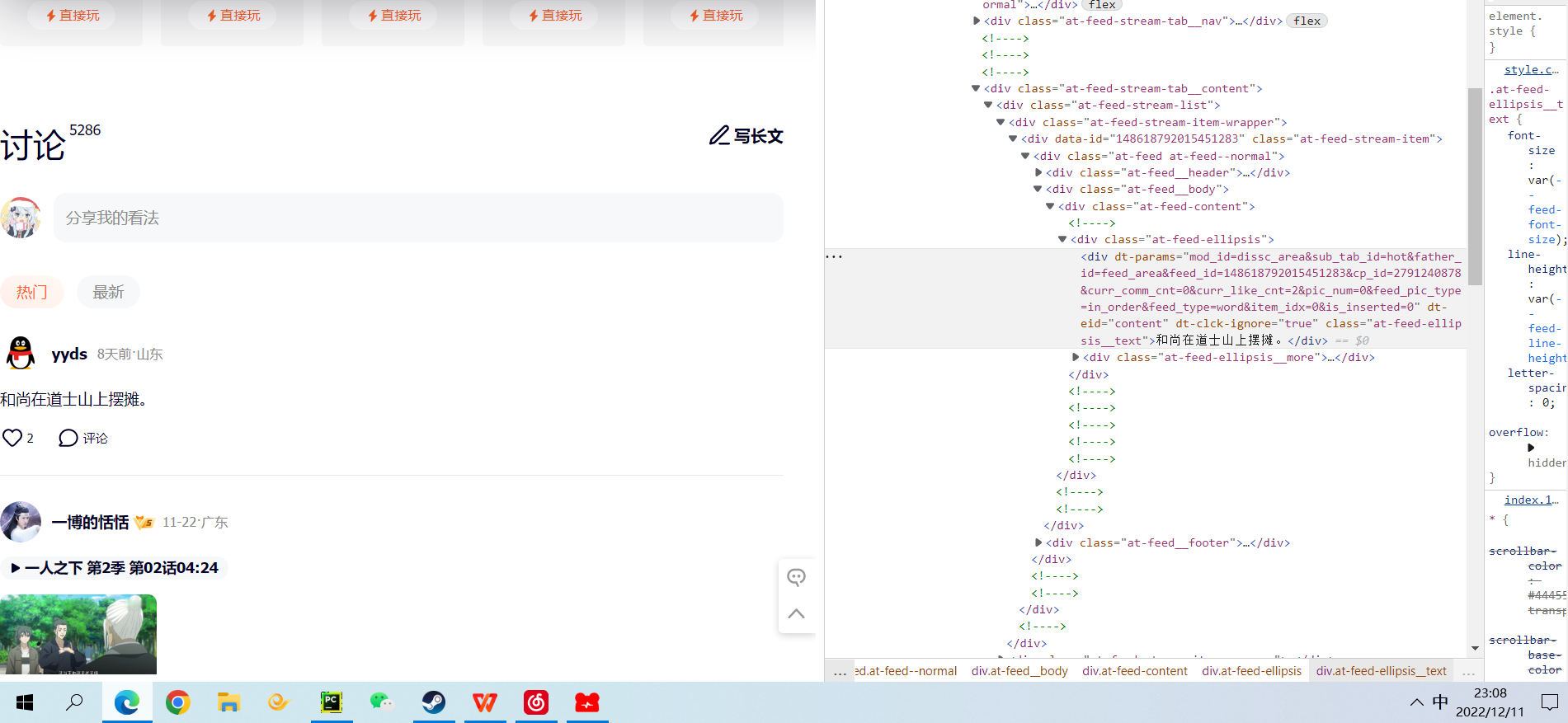
1. 作品名
2. 作品介绍
3. 超链接
4. 用户名
5. 评论时间
6. 评论内容
7. 目标网站设计原理分析

本次内容将围绕腾讯视频的动漫区展开，腾讯视频使用了大量反爬手段，包括但不限于用户每次刷新时更换目标标签位置

这是第一个程序爬取的页面，由于会随时间更新视频，所以图片仅作参考，该页面采取的是一个静态页面，因此爬取相对简单



这是第二个程序爬取的页面，由于会随时间更新视频，所以仅作参考，该页面采取的是一个动态以及有各类反爬的页面，因此爬取相对困难，这边在刚开始爬取时遇到大量困难：定位到标签后刷新页面该标签便失效无法使用，因为反爬技术将该标签下的内容转移到另外一个标签上去

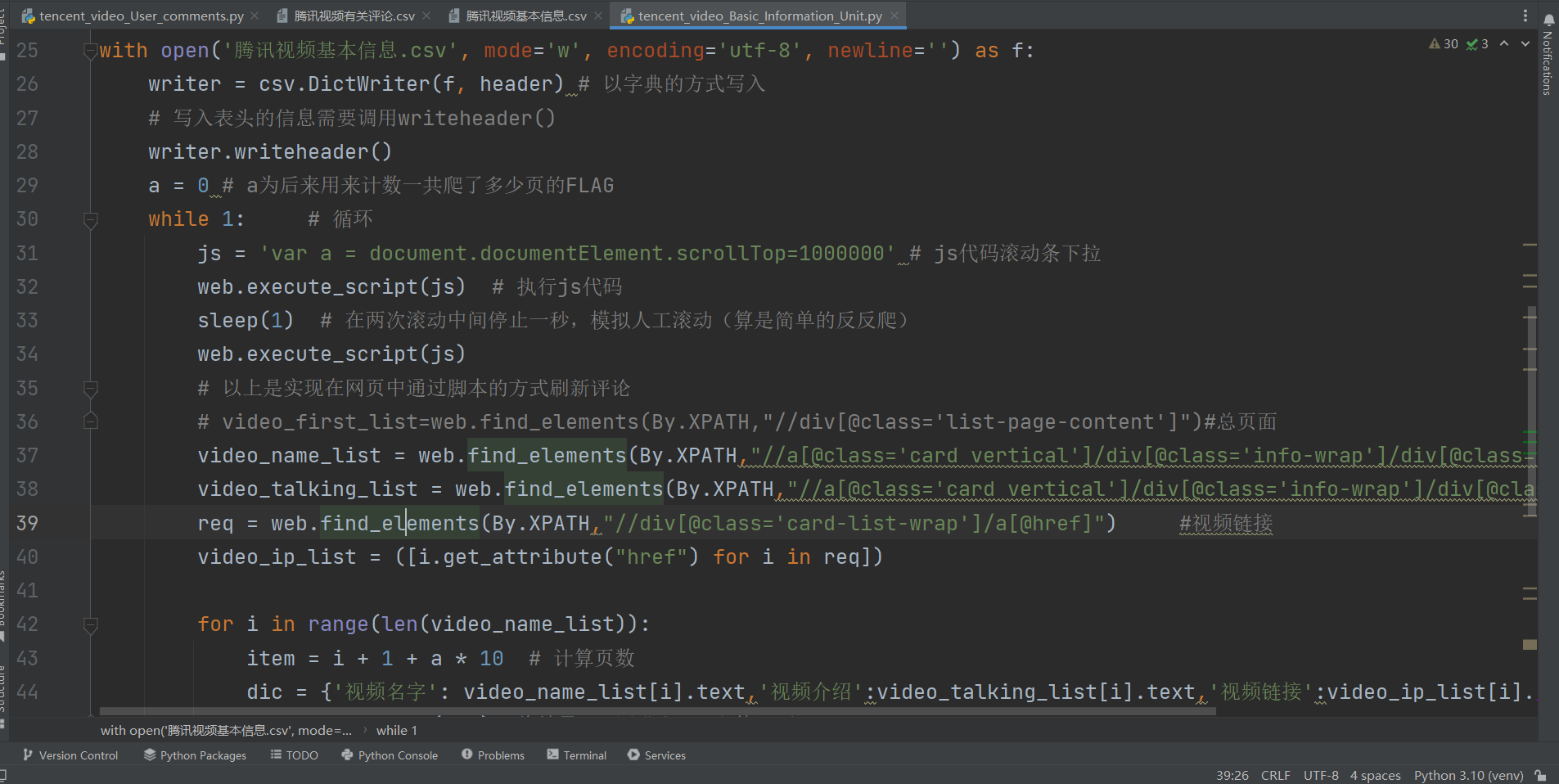


最终本程序使用的方法是使用插件将真正的标签位置将其固定住，并爬取了第一个程序爬取的作品超链接中前十条评论并存储

1. 数据采集流程与步骤说明
2. 分析网页的类型
3. 检查网页的爬虫协议
4. 对有使用反爬技术的页面进行进一步分析，并且确定框架，确定好睡眠时间，避免访问过多被系统检测到
5. 获取爬取页面下目标内容的标签，如是多页面则应该先存储后循环，避免过于混乱
6. 爬取完毕，查看结果是否与本次目标对应
7. 数据采集实现过程

本次数据采集使用到了selenium框架，并利用edge浏览器的插件，在每次爬取之前先休眠5秒让其先加载以及避免被检测到;

而后写下文件表头并用with open()函数创建文件并且写入表头，使用while函数循环使用js技术模拟鼠标下拉以达到简易的反反爬操作，而后获取内容标签，并循环获取，将其存入文件内并关闭浏览器退出，即可完成操作

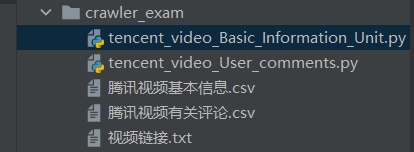


这是程序部分代码

1. 数据采集结果说明

本次两个程序爬取文件格式为csv格式，并为了爬取方便，额外写入了一个视频的超链接文件

这是运行后的采集结果，由于网站内容会更新所以仅作参考



tencent\_video\_Basic\_Information\_Unit.py程序负责爬取分区视频内容（即作品名，介绍，超链接），并将其存储在腾讯视频基本信息.csv文件中

tencent\_video\_User\_comments.py程序负责将第一个程序爬取出来的作品超链接进行循环，并爬取每一个视频前十条评论基本信息(用户名，时间，评论），并将其存储在腾讯视频有关评论.csv中

1. 任务总结与个人心得

要说的很多，首先我的爬虫作业踩在了巨人的肩膀上，爬取的程序结合了本学期的作业并独立（确信）完成了本次任务，并给予教导其他人在爬取过程中出现问题的解决方法，学会了使用油猴黑科技插件直接找到网页元素标签地址，这让我以后爬取文件更加方便，并且学到了一丢丢js内容。。。以后没饭吃就去写脚本赚钱力